



**POLITECNICO**  
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE  
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

## Two-population test for unlabelled networks

LAUREA MAGISTRALE IN MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

**Author:** DANIELE MULÈ, 945667

**Advisor:** SIMONE VANTINI

**Co-advisor:** ANNA CALISSANO, AYMERIC STAMM

**Academic year:** 2021-2022

### 1. Introduction

Networks are strongly non-Euclidean data object and it is challenging to study differences in distributions between two populations of networks relying on hypothesis testing.

As reported by [2], the traditional approach reformulates the problem in a multivariate data analysis, considering a vector of graph summary indices for each observation that, however, may not properly represent the graph structure in its complexity. Other existing approaches focus only on labelled networks, i.e., graphs that share the same exact nodes. Still, in the real world, there exist networks related to the same phenomenon that are not defined on the same set of nodes. Indeed, they may present different number of nodes or inconsistent node labels. These graphs are called unlabelled networks. Making inference on them is not trivial [1] and this also applies to tests for populations of unlabelled networks.

In this work, we introduce a general method, in which we consider a two-sample hypothesis testing approach between two populations of unlabelled networks. Our method applies to any possible type or structure of networks.

The novelty of our approach is based on the integration of the Graph Space framework [1] in network-valued hypothesis testing, in order to properly handle unlabelled networks. As in [2], we rely on permutation tests with test statistics based on inter-point distances between networks. We consider intrinsic metrics in the Graph Space and, for this reason, we refer to our method as Intrinsic Approach.

The Intrinsic Approach is implemented in the R package `nevada`.

### 2. Statistical framework

A network is a data structure defined by a set of nodes and a set of edges that specifies the connections among nodes. A network with  $n$  nodes is fully represented by its  $n \times n$  adjacency matrix  $A$ . Then, Networks are represented in the space  $X = \mathbb{R}^{n^2}$  of flattened adjacency matrices.

Since unlabelled networks may have different labels, it is not trivial to spot similarities or differences between them. As cited in [1], this problem is softened by implicit or explicit matching of nodes between two networks. Looking for a potential matching among nodes means finding an optimal alignment between networks. In the space  $X$ , this is equivalent to find the optimal permutation of networks nodes. Node permutations, that can be grouped in a set  $T$ , are possible in  $X$  via permutation matrices.

The Graph Space  $X_T := X/T$  is the quotient space of the flattened adjacency matrices space in which two networks are equivalent if they differ only by a node relabelling. In other words,  $[x] \in X_T$  includes all the networks in  $X$  that can be obtained from  $x \in X$  by permutations of the nodes labels.

A metric  $d_X$  in  $X$  implies a metric  $d_{X_T}$  in  $X_T$ :

$$d_{X_T}([x_1], [x_2]) := \min_{t \in T} d_X(tx_1, x_2) \quad (1)$$

To perform tests on unlabelled networks, we consider the permutation test, as in [2]. The permutation test

is an exact and consistent non-parametric hypothesis testing procedure. This strategy is computationally heavier with respect to parametric tests, but it does not require tight distributional assumptions or large sample sizes.

### 3. Intrinsic Approach

Given two sample of unlabelled networks in the Graph Space with sample sizes  $m_1$  and  $m_2$ , we want to test differences between their distributions.

In particular, consider a sample of i.i.d. network random variables  $G_{11}, \dots, G_{1m_1} \sim F_1$  and a sample of i.i.d. random variables  $G_{21}, \dots, G_{2m_2} \sim F_2$ ; then, the test can be expressed as:

$$H_0 : F_1 = F_2 \text{ against } H_1 : F_1 \neq F_2 .$$

As said before, we rely on the permutation test.

The most natural choice of test statistic would be a distance between the sample Fréchet means but, since the sample Fréchet mean may be not unique in the Graph Space [1], it would lead to ambiguous results. Following [2], we adopt in our method inter-point statistics based on distances between the pooled observations.

In (1), we explain that a distance in the Graph Space  $X_T$  is defined by a distance in the original space  $X$  of flattened adjacency matrices. In particular, whenever we need to compute in  $X_T$  an intrinsic distance between two networks, we are computing a distance between classes of equivalence. In  $X$ , it is the same as computing the distance between the first network and the second one subject to a node relabelling that minimizes this distance. In other words, to compute an intrinsic distance in  $X_T$  it is required to find the optimal alignment between two networks with respect to a certain distance in  $X$ . Since our method is based on intrinsic distances in  $X_T$  that require alignments between networks, we refer to our method as the Intrinsic Approach.

In the following, we summarize steps of our method. The first step requires to add null nodes to networks in the two populations that present fewer nodes. In this way, we can work with fixed-size adjacency matrices.

Then, in order to evaluate the test statistic, we compute all the inter-point distances between networks in  $X_T$  as in (1), looking for optimal nodes matching. The exact matching of nodes is practically unfeasible, since it scales rapidly with the number of nodes. Hence, we rely on approximate matching algorithms. For example, [3] propose the graph matching algorithm with indefinite relaxation of the objective function, counting on the Frank-Wolfe methodology. This algorithm is based on the Frobenius distance between networks, that is the most natural choice among distances, since it corresponds to the Euclidean distance in the  $X$ .

After distances computation, the permutation test is performed.

Finally, the Dwass-Phipson-Smyth p-value is computed.

### 4. Simulation studies

We perform simulations in order to show: (i) the effectiveness of the Intrinsic Approach, (ii) the proper test statistic to choose whenever the Frobenius distance is selected and (iii) the reliability of our method with respect to the alternative one, that we call Extrinsic Approach.

To achieve these results, we compute the Monte-Carlo estimate of the power in various scenarios.

For the Intrinsic Approach we pick the Frobenius distance, using the indefinite relaxation approximate matching algorithm.

The Extrinsic Approach share the same statistical framework related to the Graph Space proposed for the Intrinsic Approach, but it employs extrinsic distances in  $X_T$  based on maps from  $X_T$  to a lower dimensional space that are invariant to node relabelling. For this reason, the Extrinsic Approach does not require any alignment between networks. In simulations we choose the spectral distance for the Extrinsic Approach.

In the first simulation, we compare different inter-point test statistics. We replicate the same simulation study led by [2], concentrating on IP-StudentFisher, energy, density-based and generalized edge-count statistics. In particular, we generate two samples of unlabelled networks whose edge weights follow an i.i.d. binomial distribution. We treat these graphs as unlabelled networks in the Graph Space, i.e., as if we do not have any a priori knowledge of correspondence in nodes labels. We consider 3 scenarios in which the distributions of the two samples differ by location-only, scale-only and both location-scale differences, respectively.

We notice that our method gives expected results. Both for location-only and location-scale differences, the power grows coherently with the increasing value of differences; this is also valid for scale-only, except for the energy statistic that has low power. The IP-StudentFisher statistic performs well in each situation, and it is the best statistic in detecting differences in mean and simultaneously in mean and variance; the density-based statistic has similar performance and it is slightly more powerful than the IP-StudentFisher in finding scale-only differences. The generalized edge-count has lower power in each scenario, meanwhile the energy statistic reaches good performance in location-only and location-scale differences but, as reported before, it has very low power in detecting differences in variance.

In the second simulation, we compare the Intrinsic Approach with the Extrinsic Approach. Espe-

cially, we show that our method is able to detect actual differences in distribution of the two samples in situations in which the alternative one fails. More precisely, the Intrinsic Approach is not critically restricted to the choice of a specific metric, since intrinsic distances in the Graph Space preserve the original structure of the graph. Instead, the choice of a distance for the Extrinsic Approach is crucial, since it maps the network in a lower dimensional space with resulting loss of information; as a matter of fact, the chosen embedding may not capture important features of the graph structure.

We generate networks that share the same spectrum distribution, but two different sets of eigenvectors, one for the first sample and the other for the second sample.

To make a fair comparison, we choose the IP-StudentFisher statistic, since it reaches good performance for both methods.

As expected, the Extrinsic Approach with spectral distance is not able to detect differences between samples under the alternative hypothesis of different distributions, meanwhile the Intrinsic Approach has very high power.

## 5. Application to football data

To show the usefulness of our method, we apply the Intrinsic Approach with Frobenius distance and the IP-StudentFisher statistic on a real example in the field of sports analytics.

Our aim is to analyse football passing strategies. This task can be achieved by the study of the players' passing network, i.e., a weighted graph with nodes representing the players in a team and edges expressing the number of passes between two players in a match. Football passing networks are perfect examples of unlabelled networks since, even in the same team, there may be different players that take part to games.

We create passing networks from a football database available in StatsBomb Open Data repository [4]. We focus on the UEFA EURO 2020 and LaLiga competitions.

In order to understand relations between samples, we perform a hierarchical agglomerative clustering. Moreover, Fréchet means are computed, to make qualitative comments on clusters.

In UEFA EURO 2020 competition, we consider two-sample tests in which we compare two by two passing strategy distributions between all the 24 competing teams in UEFA EURO 2020. We assign to each team a passing network for every game it plays. We notice that teams can be grouped in 4 clusters of passing networks. Qualitatively, Fréchet means tell us that clusters are based on high, medium-high, medium-low and low number of passes during the match.

In LaLiga competition, we compare the passing strategies adopted by different Barcelona managers

from October 2004 to May 2021, i.e.: Frank Rijkaard, Pep Guardiola, Tito Vilanova, Gerardo Martino, Luis Enrique, Ernesto Valverde, Quique Setién and Ronald Koeman. We assign to each manager a passing network for every match they play. We notice that Rijkaard implemented a different passing strategy with respect to all the other managers and, qualitatively looking at Fréchet means, it is probably because large number of passes are restricted only to few players in comparison to other managers'. Observing Guardiola, Vilanova and Martino Fréchet means, we can deduce that these managers adopted a passing strategy that favours large number of passes between all the players. This latter comment is applicable in a more modest way to the other managers, i.e., Setién, Koeman, Enrique and Valverde that can be grouped as an unique class.

## 6. Discussion

The Intrinsic Approach is a new method that allows hypothesis testing between two populations of unlabelled networks. Its innovation lies in the integration of Graph Space theory in testing framework. The strength of our method is supported by a great flexibility in the choice of test statistics and intrinsic distances. In contrast to the alternative method, i.e., the Extrinsic Approach, the Intrinsic Approach preserves all the information about networks, reaching reliable results. Moreover, the permutation tests do not need critical distributional assumptions, nor particular requirements on the sample sizes.

The main limitation of this approach derive directly from drawbacks of the Graph Space and it is related to the fact that approximate matching algorithms slow down the entire method.

However, new algorithms may be studied and developed in order to improve this aspect and to support other intrinsic distances besides the Frobenius one. Moreover, further developments can be done providing quantitative tools to properly interpret results from testing in populations of unlabelled networks.

## References

- [1] Anna Calissano, Aasa Feragen, and Simone Vantini. Populations of unlabeled networks: Graph space geometry and geodesic principal components. *MOX Report*, 2020.
- [2] Ilenia Lovato, Alessia Pini, Aymeric Stamm, and Simone Vantini. Model-free two-sample test for network-valued data. *Computational Statistics & Data Analysis*, 144:106896, 2020.
- [3] Zihuan Qiao and Daniel Sussman. iGraphMatch: an R Package for the Analysis of Graph Matching, 2021.
- [4] StatsBomb Services Limited. StatsBomb Open

Data, 2022. URL: <https://github.com/statsbomb/open-data>.